4C

УДК 004.934

М.М. Сажок, Р.А. Селюх, О.А. Юхименко
Міжнародний науково-навчальний центр інформаційних технологій та систем,
м. Київ, Україна
mykola@uasoiro.org.ua, selyukh@uasoiro.org.ua, yukhymenko@uasoiro.org.ua

Адаптація акустичних моделей фонем до голосу диктора для пофонемного розпізнавання ізольованих слів української мови

У статті розглядаються проблеми адаптації моделей фонем до голосу диктора для пофонемного розпізнавання ізольованих слів української мови. Описується метод адаптації під назвою «лінійна регресія максимальної правдоподібності» (MLLR). У рамках цього методу шляхом оптимізації значення критерію розпізнавання отримуються матриці лінійних перетворень, за якими адаптуються акустичні моделі фонем. Наводяться результати експериментальних досліджень розпізнавання мовлення адаптованої системи. Аналізуються дані розпізнавання адаптованих моделей у залежності від кількості слів в адаптаційній вибірці.

Вступ

Пофонемне розпізнавання мовленнєвого сигналу передбачає формування усномовного паспорта диктора, що включає акустичні моделі фонем [1]. Оцінка параметрів моделей фонем проводиться за навчальною вибіркою, яка повинна містити все фонемне ромаїття мови. Досвід формування таких вибірок показав, що їх обсяги повинні бути настільки великими, що диктору необхідно витратити багато годин для запису мовлення, щоб досягти прийнятної надійності при пофонемному розпізнаванні ізольованих слів з великих словників [2]. За таких умов використання технологій розпізнавання усної мови суттєво обмежується. Чи можна скоротити обсяг вибірки, потрібної для настроювання на голос диктора? Щоб дати ствердну відповідь на це питання, розглядається ще один клас задач мовленнєвої інформатики: задача адаптації на голос диктора. Ця задача передбачає попереднє проведення навчання розпізнаванню на голос деякого опорного диктора або кооперативу дикторів. Потім здійснюється коригування параметрів акустичних моделей фонем для нового диктора на відносно невеликій вибірці – адаптаційній. Також адаптація може проводитися і до зміни умов розпізнавання, як, наприклад, перехід на інший канал отримання усномовної інформації (інший мікрофон, телефонна лінія).

Метою роботи є дослідити та застосувати до українського мовлення один із найбільш поширених підходів до адаптації на голос диктора при пофонемному розпізнаванні окремо вимовлюваних слів.

У попередніх дослідженнях з адаптації на голос диктора проводилося коригування акустичних генеративних моделей цілих слів [3]. На теперішньому етапі ми переходимо до пофонемного розпізнавання.

1. Постановка задачі адаптації та шляхи її вирішення

Нехай маємо оцінені параметри акустичних генеративних моделей фонем на підставі ітераційних процедур для опорного диктора або для кооперативу дикторів [3], [4]. Зокрема для кожної з трьох фаз-станів фонеми φ (рис. 1) нам відомі вектор математичного сподівання $\mu = [\mu_1, \mu_2, ..., \mu_n]^T$ та коваріаційна матриця Σ, розмірністю $n \times n$, де n – розмірність вектора первинних ознак сигналу.

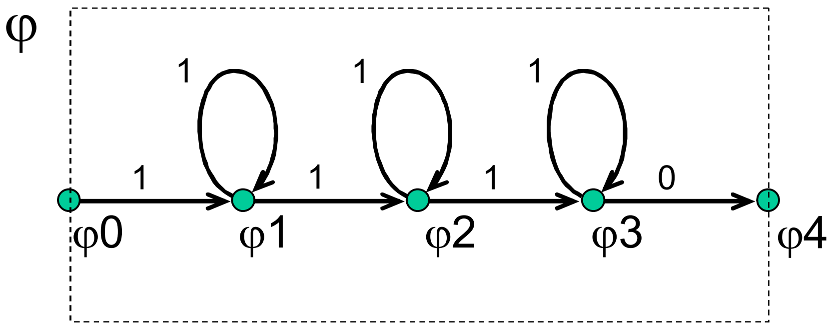


Рисунок 1 – Генеративна модель фонеми φ з трьома фазама-станами φ_1, φ_2, φ_3. Додаткові неемітентні стани φ_0 і φ_4 вводяться для сполучення з іншими моделями фонем. Число поруч із дужкою вказує на кількість часових відліків, за які здійснюється перехід

Припускається, що існує лінійне перетворення, яке переводить початкові вектори математичного сподівання у вектори математичного сподівання для нового диктора. Ефектом цього перетворення є зсув середніх значень параметрів моделей фонем та зміна дисперсій цих параметрів у початковій системі таким чином, що кожний стан у системі акустичних моделей фонем може точніше генерувати дані адаптації.

Лінійне перетворення для середніх значень записується у вигляді:

$$\hat{\mu} = W\xi , \qquad (1)$$

де $\hat{\mu}$ – вектор матсподівання нового диктора, W є матрицею розмірністю $n \times (n + 1)$, ξ – вектор розширеного матсподівання

$$\xi = \left[w, \mu_1, \mu_2, ..., \mu_n\right]^T , \qquad (2)$$

де w представляє нев'язку, початкове значення якої фіксоване і дорівнює 1.

У свою чергу, матриця W розкладається на добуток

$$\mathrm{W} = [\ \mathrm{b}\ \mathrm{A}\ ] , \qquad (3)$$

де A є матрицею лінійних перетворень розмірністю $n \times n$, а b представляє вектор ухилу.

В такій формі перетворення зручніше обчислюється в умовах неперервного розподілу за нормальним законом.

Перетворення коваріаційних матриць не досліджувалося, тому його опис пропускаємо.

Матриці лінійних перетворень отримуються шляхом оптимізації значення критерію розпізнавання. Одним з таких оптимізаційних алгоритмів є лінійна регресія максимальної правдоподібності (Maximum Likelihood Linear Regression – MLLR) [4]. Стани фонеми автоматично поділяються на певну кількість класів регресії методами векторного квантування, а потім для кожного класу регресії оцінюється своя матриця лінійних перетворень за ітераційною процедурою.

Ця ж процедура використовується і у випадку апроксимації фаз-станів фонем сумішшю нормальних законів – ґаусіанів. Тоді до класів регресії входять окремі ґаусіани.

2. База даних і знань

У дослідженнях ми використали україномовний багатодикторний мовленнєвий корпус, який містить понад 30 000 реалізацій слів і тисячі речень близько 100 дикторів, що мешкають у різних областях України. Реалізації слів зберігають частотні пропорції фонем і є фонетично збалансованими, при підборі слів також враховувалися їх частотні характеристики [2]. Цей мовленнєвий корпус було створено завдяки гранту Президента України для обдарованої молоді, контракт № 32 від 30.05.2006 р.

Взято до розгляду матеріал з мовленнєвого корпусу, записаний з голосу 62 дикторів. Цю основну вибірку розділено на дві частини. Перша частина (49 дикторів) призначена для використання як навчальна вибірка.

Друга частина вибірки (14 дикторів) має такі властивості: (1) набір з 241 слова, вимовлений кожним диктором був один і той же; (2) ніяке слово з другої частини вибірки не вимовлялося жодним диктором з першої вибірки. Ця частина вибірки призначена як для адаптації, так і для контролю. Завдяки властивостям другої частини вибірки ми маємо змогу проводити адаптацію для різних дикторів на одному і тому ж наборі слів, а також виключити перетинання слів з контрольної та навчальної вибірок.

При розпізнаванні використовувався словник обсягом 2170 слів, який включав усі слова з основної вибірки.

3. Експериментальні дослідження адаптації

Було проведено початкове оцінювання параметрів акустичних моделей фонем у мел-кепстральному просторі ознак, доповненому дельта-коефіцієнтами та «прискоренням», на навчальній вибірці, описаній у попередньому розділі. Кожна фаза-стан фонеми з алфавіту фонем української мови моделювалася сумішшю нормальних законів, кількість яких варіювалася для кожної серії експериментів від 8 до 16.

Адаптація проводилася для кожного диктора на різній кількості реалізацій слів, узятих з другої частини вибірки. При адаптації отримали 13 класів регресії, для кожного класу було оцінено свою матрицю перетворення.

Розпізнавання проводилося для кожного диктора окремо на адаптованих для нього моделях. Усереднену надійність розпізнавання для всіх 14 дикторів подано на рис. 2 для двох серій експериментів. Очевидно, кращі результати показали моделі з 16 ґаусіанами. Для цих моделей розраховано надійність розпізнавання окремо для кожного диктора.

Результати, наведені в табл. 1, показують, що після адаптації на голос нового диктора надійність розпізнавання в середньому виросла на 3,03% для адаптаційної вибірки обсягом у 30 слів, на 3,82% – для 60 слів, на 4,64% – для 100 слів, на 5,55% – для 150 слів.

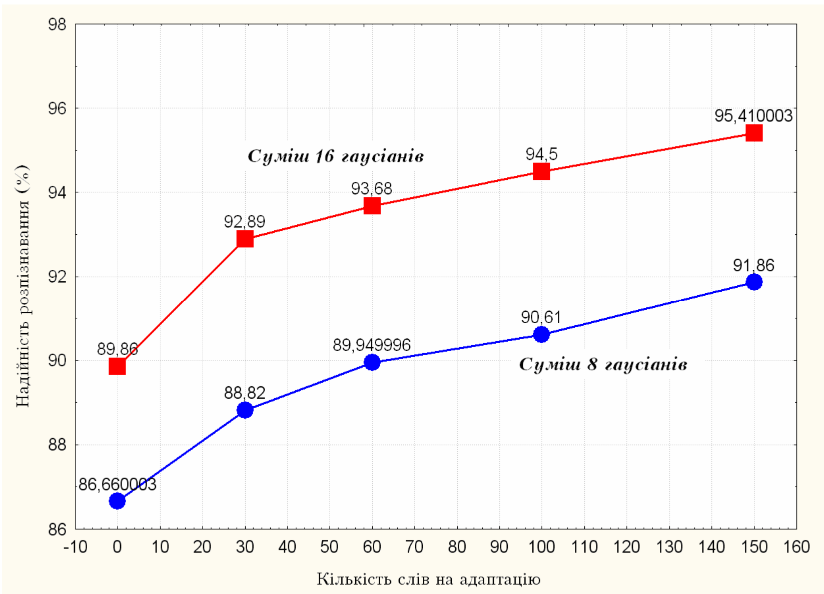


Рисунок 2 – Усереднена надійність розпізнавання для серій експериментів з різною кількістю ґаусіанів

Висновки

Експериментально підтверджено, що адаптація на голос диктора є перспективною технологією. Користувачеві достатньо вимовити лише декілька десятків слів українською мовою, щоб отримати прийнятну надійність розпізнавання великих словників. Вперше отримані результати адаптації для ізольованих слів української мови, які відповідають рівню європейських досліджень [5].

Подальші роботи будуть спрямовані на підвищення якості адапації, зокрема шляхом перетворення матриць дисперсії та залучення до розпізнавання оцінки довжини голосового тракту диктора. Будуть також досліджені інші простори первинних ознак сигналу. Планується працювати не лише з ізольованими словами, а й зі злитим мовленням, збільшити обсяги словника.

Таблиця 1 – Надійність розпізнавання (%) для групи нових дикторів до і після адаптації на різну кількість слів. Кількість сумішей ґаусіанів у моделях фонем – 16

Кількість слів на адаптацію / **Диктори**	**0** (без адаптації)	**30**	**60**	**100**	**150**
1. Анна	93.78	95.74	96.32	95.88	97.43
2. Богдан	80.50	88.90	89.87	91.06	93.77
3. Валентина	95.02	95.39	96.13	96.17	94.50
4. Ганна	91.29	92.28	91.92	92.48	93.04
5. Дмитро	92.12	95.40	96.60	98.01	97.07
6. Катерина	79.25	84.90	85.91	88.37	91.57
7. Олена	90.46	93.23	94.75	95.32	96.70
8. Олеся	92.53	93.23	94.75	95.32	96.70
9. Руслан	89.21	92.96	94.48	94.75	94.87
10. Сергій	95.81	96.55	96.60	97.16	97.43
11. Слава	89.21	90.93	91.35	92.06	92.68
12. Тетяна	87.14	91.34	92.64	94.33	97.44
13. Юрій	89.21	93.70	94.20	96.03	95.60
14. Юрій 2	92.53	95.94	95.95	96.03	97.07
В середньому по групі	89.86	92.89	93.68	94.50	95.41

Література

1. Vintsiuk Taras. Speaker Voice Passport for a Spoken Dialogue System / Taras Vintsiuk, Mykola Sazhok // Proceedings of the 3rd International Workshop «Speech and Computer» – Specom'98. – St.-Petersburg, 1998. – P. 275-278.
2. Vasylieva N. Text Selection for Training Procedures under Phoneme Units Variety / N. Vasylieva, M. Sazhok // Proceedings of the 10th International Conference on Speech and Computer – SpeCom'2005. – Patras, 2005, P. 69-76.
3. Винцюк Т.К. Анализ, распознавание и смысловая интерпретация речевых сигналов / Т.К. Винцюк. – Киев : Наукова думка, 1987.
4. Young S.J. HTK Book, version 3.1 / Young S.J. [et al]. – Cambridge University, 2002. – 355 p.
5. Červa P. Map Based Speaker Adaptation in Large Vocabulary Speech Recognition of Czech Language / P. Červa, J. Nouza // Proc. of Radioelektronika 2004, (April 2004, Bratislava). – Slovak Republic. – P. 108-111.

М.М. Сажок, Р.А. Селюх, О.А. Юхименко

Адаптация акустических моделей фонем на голос диктора для пофонемного распознавания изолированных слов украинского языка

В статье рассматриваются проблемы адаптации моделей фонем на голос диктора для пофонемного распознавания изолированных слов украинского языка. Описывается метод адаптации под названием «линейная регрессия максимального правдоподобия» (MLLR). В рамках этого метода путем оптимизации значения критерия распознавания получаем матрицы линейных преобразований, по которым адаптируются акустические модели фонем. Приводятся результаты экспериментальных исследований распознавания речи адаптированной системы. Анализируются данные распознавания адаптированных моделей на разном количестве слов.

Стаття надійшла до редакції 09.06.2009.